DOCUMENT RESUME

ED 111 844 TM 004 811

AUTHOR Besel, Ronald
TITLE A Comparison of Emrick and Adam's Mastery-Learning Test Model with Kriewall's Criterion-Referenced Test Model.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
REPORT NO SWRL-TM-5-71-04
PUB DATE 21 Apr 71
NOTE 17p.; For related documents, see TM 004 812 and 814

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Bayesian Statistics; *Comparative Analysis; *Criterion Referenced Tests; Cutting Scores; Decision Making; Educational Diagnosis; Grouping (Instructional Purposes); *Mathematical Models; Probability; Psychometrics; Test Interpretation
IDENTIFIERS Kriewalls Criterion Referenced Test Model; Mastery Learning Test Model (Emrick and Adams); *Mastery Tests

ABSTRACT

The assumptions of the Criterion-Referenced Test (CRT) model proposed by Kriewall are compared to those of Emrick and Adam's Mastery-Learning (ML) model. Testing, in the context of instructional management, serves three general purposes: performance evaluation (achievement of objectives), placement (classification of students for instruction), and diagnosis of learning deficiencies. Both of the test models discussed here assess the achievement of objectives; they differ in the types of objectives for which they are best suited. Both test models have potential usefulness for making placement decisions, but only the ML model is likely to be useful in diagnosing learning deficiencies. The applicability of each model for instructional management decisions is discussed. (Author/DEP)

TM

SWRL

SOUTHWEST REGIONAL LABORATORY
TECHNICAL MEMORANDUM

DATE: April 21, 1971

NO: TM 5-71-04

TITLE: A COMPARISON OF EMRICK AND ADAM'S MASTERY-LEARNING TEST MODEL WITH KRIEWALL'S CRITERION-REFERENCED TEST MODEL

AUTHOR: Ronald Besel

ABSTRACT

The assumptions of the Criterion-Referenced Test (CRT) model proposed by Kriewall are compared to those of Emrick and Adam's Mastery-Learning (ML) model. The applicability of each model for instructional management decisions is discussed.

THE MASTERY-LEARNING TEST MODEL: COMPARISON WITH KRIEWALL'S CRT MODEL

Numerous psychometric models have been used for interpreting testing data (Lord and Novick, 1968); most of these models are appropriate for norm-referenced tests (NRT). Kriewall (1969) criticized the use of NRT models for interpreting criterion test data. He proposed a model for criterion tests which he called the CRT model. In the same critical spirit, Emrick and Adams (1970) proposed a Bayesian mastery-learning (ML) model.

Testing, in the context of instructional management, serves three general purposes: performance evaluation (achievement of objectives), placement (classification of students for instruction), and diagnosis of learning deficiencies. Both of the test models discussed here assess the achievement of objectives; they differ in the types of objectives for which they are best suited. Both test models have potential usefulness for making placement decisions but only the ML model is likely to be useful in diagnosing learning deficiencies.

Psychometric Assumptions

The ML model assumes that a test measures a single skill and that there are only two true states of proficiency with respect to that skill. Each individual tested is in either the mastery (M) or non-mastery ($\overline{M}$) state at the time of testing. The CRT model assumes, likewise, that a test is a measure of a single skill - - defined by a specified content objective (SCO), i.e., a rule or procedure for generating a class of problems - - but, proficiency is assumed to be a continuum between mastery and non-mastery.

Both the CRT and the ML models assume that an individual's responses to the separate items on a test can be treated as a sequence of independent Bernoulli trials, each having the same probability of success. The consequences of this assumption are:

(a) The probability that a given individual will give a correct response to any item from the test (or from the domain from which the items were selected) is the same for all items;

(b) No learning occurs during the time of test administration;

(c) The outcome of any trial (item response) is independent of the outcomes of every other trial.

Let,

x_{ia} represent the response of individual a to item i (trial i)

$$x_{ia} = \begin{cases} 1 \text{ if the correct response is given} \\ 0 \text{ if an incorrect response is given} \end{cases} \quad (1)$$

X_a = observed score (number of correct responses) for individual a

$$X_a = x_{1a} + x_{2a} + \ldots + x_{na} \quad (2)$$

The subscript "a" will be deleted when the referent for an observed item response or test score is not a particular individual. Since both models assume that test performance can be represented mathematically as a sequence of independent Bernoulli trials, each hypothesizes that if an individual is repeatedly given parallel tests, his score distribution will be binomial with the probability of a correct item response (p_a) as the distribution parameter. For a n-item test, the score distribution function is:

$$P(X_a) = \binom{n}{X_a} p_a^{X_a} (1-p_a)^{n-X_a} \quad (3)$$

where,

$$\binom{n}{X_a} = \frac{n!}{X_a!\ (n-X_a)!} \qquad (4)$$

The CRT and ML models differ in their interpretation of the parameter p_a. The CRT model assumes that p_a is equivalent to a "true score" as in typical NRT models. For any individual, "true proficiency" is estimated from observed score.

$$\hat{p}_a = \frac{X_a}{n.} \qquad (5)$$

The ML model assumes that p_a is a single constant value $(1-\beta)$ for all individuals in the mastery (M) state and a constant α for all individuals in the non-mastery ($\bar{M}$) state:

α = The probability that an individual in the $\bar{M}$ state will give a correct item response.

β = The probability that an individual in the M state will give an incorrect item response.

Both α and β are assumed to have true values which are characteristic of the test. They must be estimated from the responses of some reference group of individuals. Two conditional distributions can represent the expected score distributions for all individuals when the ML model is employed:

$$P(X/M) = \binom{n}{X} (1-\beta)^X \beta^{n-X} \qquad (6)$$

$$P(X/\bar{M}) = \binom{n}{X} \alpha^X (1-\alpha)^{n-X} \qquad (7)$$

The CRT model characterizes each individual tested by his estimated "true proficiency" using equation (5). The ML model, on the other hand,

characterizes each individual by his estimated probability of being in the mastery state. This probability, can be computed using Bayes formula:

$$P(M/X) = \frac{PR(M) \cdot P(X/M)}{P(X)} \qquad (8)$$

where PR(M) equals the prior probability that the individual is in the mastery state and,

$$P(X) = PR(M) \cdot P(X/M) + PR(\bar{M}) \cdot P(X/\bar{M}) \qquad (9)$$

where, $PR(\bar{M}) = 1-PR(M)$ (10)

The ML model requires valid procedures for estimating prior probabilities. Procedures applicable to SWRL instructional programs will be discussed in a later section.

Kriewall assumes that rigorous item-sampling procedures will be followed to construct parallel criterion-referenced tests. Emrick and Adams do not specify a test construction procedure for a ML test; Kriewall's method would be applicable but the ML model may also be valid for criterion tests constructed using less rigorous procedures. The reference group used to estimate the α and β parameters could also be employed to test the equal item-difficulty assumption inherent in both models. If this assumption is found to be empirically untennable, the test model or the SCO domain may require modification.

Measurement errors are interpreted differently by the two test models. Kriewall assumes that all measurement errors are random with an expected value equal to zero. The observed score can then be interpreted as an unbiased estimate of the percentage of the items in the content domain for which the individual knows the correct response. The interpretation of a test which is biased (e.g., a multiple-choice

test) is not discussed by Kriewall, the CRT test model has an intuitively appealing intrepretation only for unbiased tests. The α and β parameters of the ML model represent two types of bias errors. Both constructed-response and selected-response tests can be interpreted readily when the ML model is employed. The α value for a selected-response test is likely to be significantly larger than the α value for a comparable constructed-response test. Selected-responses tests, however, may achieve smaller β parameters.

Prior Probabilities

A Bayesian model is, in general, more efficient than one employing classical statistics to the extent that prior probabilities can be precisely estimated. Two general classes of prior probability estimates can be used by the ML model. The first class includes estimates of the proportion of an appropriate reference group which is in the mastery state. In interpreting an individual's observed score, he is treated as a random sample from the reference group.

The second class of prior probability estimates includes only methods which use other past or present numerical information relevant to an individual. These will be called personalized prior probabilities and the subscript (a) will be added to the symbol for a prior probability.

$PR_a(M)$ = personalized prior probability for individual (a).

For criterion-referenced tests three types of pupil performance data seem to be most relevant to estimating personalized prior probabilities. The first is the P(M/X) value for a similar objective for which assessment was made in the recent past (e.g., skill in reading

the words on a current vocabulary list should be a potential value for $PR_a(M)$ for the vocabulary list of the next instructional unit). If a hierarchial relationship between objectives exist, another reasonable estimate of a prior probability would be the probability of mastery of the objective directly below the one in question. This approach may use either current or past test data depending on the testing schedule. A third method employs the scores on a pretest; preferably a parallel version of the posttest is used.

Instructional Decisions

Either the estimated proficiency ($\hat{p}_a$) of the CRT model or the probability of mastery, P(M/X), of the ML model can be used as a decision variable when classifying pupils for instructional purposes. In theory, the decision variable can be used to classify pupils into m subgroups where m can take any integer value which does not exceed the number of discrete levels assumed by the decision variable. In a school setting it is not likely that classification into more than three groups will be practical. Kriewall treats only the case of classification into two groups - masters and non-masters.

For the CRT model, the following steps are taken in selecting the test length (n) and the acceptable "passing" score (c) for the two-group classification.

1. A minimal acceptable proficiency (criterion level) is selected. The nominal student is defined to be one whose proficiency equals this criterion level (p_1).

2. The acceptable probability of a Type I error (classifying a nominal student as a non-mastery) is chosen.
3. A proficiency level, p_2, less than p_1 is selected to represent the "nominal non-master."
4. The acceptable probability of a Type II error (classifying a nominal non-master as a master) is chosen.
5. Values of n and c which satisfy the Types I and II error limits for the chosen p_1 and p_2 proficiencies are solved for iteratively.

For the ML model, the "passing score," c, for the two-group classification problem can be computed using a simple expected loss model.

Let,

L(c) = loss for a passing score of X = c

Type I error: classifying a master as a non-master (false fail)

Type II error: classifying a non-master as a master (false pass)

L_1 = cost of making a Type I error

L_2 = cost of making a Type II error

$E\left\{ L(c) \right\}$ = expected loss

The expected loss for any selected value c is:

$$E\left\{ L(c) \right\} = L_1 \sum_{X=0}^{c-1} P(M/X) \cdot P(X) + L_2 \sum_{X=c}^{n} P(\bar{M}/X) \cdot P(X) \qquad (11)$$

Increasing c by 1 will result in deleting one term from the second summation in equation (11) and adding one term to the first. The expected loss is minimized by including in the first summation only those terms for which $L_1 \cdot P(M/X)$ is less than $L_2 \cdot P(\bar{M}/X)$. Equation (11) assumes that a single "passing score" must be selected for classifying

each individual in the group. This restriction need only be made if each individual in the group is assumed to have the same prior probability of being in the mastery state. If personalized prior probabilities can be estimated, the computed P(M/X) for each individual can be used directly as a decision variable. The P(M/X) value which yields the minimum expected loss is then:

$$L_1 \cdot P(M/X) = L_2 \cdot P(\overline{M}/X) \tag{12}$$

$$P(M/X) = \frac{L_2}{L_1 + L_2} \tag{13}$$

This probability of mastery value is a criterion level in the same sense that "80 percent correct answers" is used as a criterion level. If the computed P(M/X) exceeds this criterion level, the individual is classified as a master; otherwise he is classified as a non-master. The ratio, L_2/L_1, will be referred to as the loss ratio. The effect of prior probability on the selection of the optimal c value is illustrated by Figure 1. For a loss ratio equal to 3, c should be set equal to 5 for prior probabilities between .15 and .59; c should equal 4 for prior probabilities between .59 and .92 and equal to 3 for prior probabilities greater than .92. Thus, if personalized prior probabilities are employed, the test will not have a fixed passing score.

Test Length and Sequential Testing

Figure 2 illustrates the ML model for a one-item test. The two sets of α, β parameters chosen are representative of typical selected response and constructed response tests.

The mathematical interpretation of the ML model permits a simplified computation of P(M/X) when one or more items are added sequentially to a n-item test. If a single item is added to a n-item test, Figure 2 can be used to obtain a revised or posterior value for P(M/X); the P(M/X) for a n-item test is used as the prior probability value for a (n + 1) -item test.

Example 1

(a) Prior probability assumed to be .6.

(b) Student gave 4 correct responses on a 5-item test.

(c) α = .5, β= .1

(d) Student gave a correct response on a sixth item.

From Figure 1, P(M/X) for the 5-item test is .76; using this value as a prior probability, P(M/X) for the 6-item test from Figure 2 is .85.

The effect of doubling the test length can be estimated from Figure 1 in a similar manner.

Example 2

(a) Prior probability assumed to be .5.

(b) Student gave 8 correct responses on a ten item test.

(c) α = .5, β= .1

Each combination of 5-item test responses which result in 8 correct responses on a ten item test are tabulated.

Initial 5-items X	P(M/X)	Final 5-items X	Combined P(M/X)
3	.190	5	.82
4	.677	4	.82
5	.950	3	.82

Note that the P(M/X) for the combined 10-item test is the same for each possible combination of correct responses on initial and final tests.

Mastery Tests vs Criterion Referenced Tests

Mastery tests may be viewed as a special case of criterion referenced tests:

(a) The criterion level is a perfect score.

(b) The only true scores are assumed to be 0 and n (for a n-item test); all intermediate scores are due to measurement error.

It is the second characteristic (b) which permits the application of a simple decision rule to determine the "passing" score.

The ML model is designed specifically for mastery tests; the CRT model is appropriate for situations where it is meaningful to speak of "degree of performance." For example, the CRT model may be used to estimate the percentage of words in a lengthy vocabulary list that a student can read. The CRT model would seem to be most appropriate for evaluating performance when the content domain of the objective is so large as to require an item-sampling procedure. The ML model can be used for single-item tests; longer tests are conceptualized as replications of single-item tests. The ML model is most appropriate for narrowly defined behavioral objectives for which performance can be conceptualized as "all or nothing."

To be useful as a diagnostic tool, a test must break down performance into separate skills for which prescriptive treatments are available and effective. The ML model is well suited to measuring skills at a level of specificity which is desirable for remedial instruction. The ML

model can also be applied to multiple choice tests which can be constructed (Guttman, 1970) so that the distractor responses represent meaningful types of learning deficiencies.

The ML model is designed to be used in making placement or classification decisions from diagnostic information. Individually tailored drill and practice exercises, tutorials, or small group instruction are types of treatments which may be prescribed from mastery-learning diagnosis. The CRT model is appropriate for making placement decisions based on degree of proficiency rather than diagnosis of learning deficiencies. Degree of proficiency may be treated as a measure of aptitude for future learning. If aptitude is viewed as the amount of time required by the learner to attain mastery of a learning task (Carroll, 1963) placement decisions based on aptitudes may improve the efficiency of instruction. The formation of instructional groups for initial instruction is an example of this type of placement decision.

Comparision of the two test models leads to the following conclusions:

1. The ML model is applicable to very short tests -- 5-items or fewer -- and is appropriate for instructional decisions related to specific behavioral objectives.
2. The CRT model is suited for longer tests -- approximately 20 or more items -- unless testing can be done sequentially by item. The CRT model may be better than the ML model for more general behavioral objectives (e.g., an extensive content domain).
3. A CRT model may be used as the basis for forming instructional groups for a group-oriented mode of instruction; mastery tests

and the ML model may be more appropriate for diagnostic testing, prescribing practice, tutoring, or other types of remedial instruction.

REFERENCES

Carroll, J.A. "A Model of School Learning," Teachers College Record, 1963, (Vol. 64), 723-733.

Emrick, J.A., & Adams, E.N. "An Evaluation Model for Individualized Instruction," Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minnesota, 1970.

Guttman, L. "Integration of Test Design and Analysis," in the Proceedings of the 1969 Invitational Conference on Testing Problems, Educational Testing Service, 1970, 53-65.

Kriewall, T.E. "Applications of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction," Technical Report No. 103, Wisconsin Research and Development Center for Cognitive Learning, Madison, Wisconsin, 1969.

Lord, F.M., & Novick, M.R. Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.

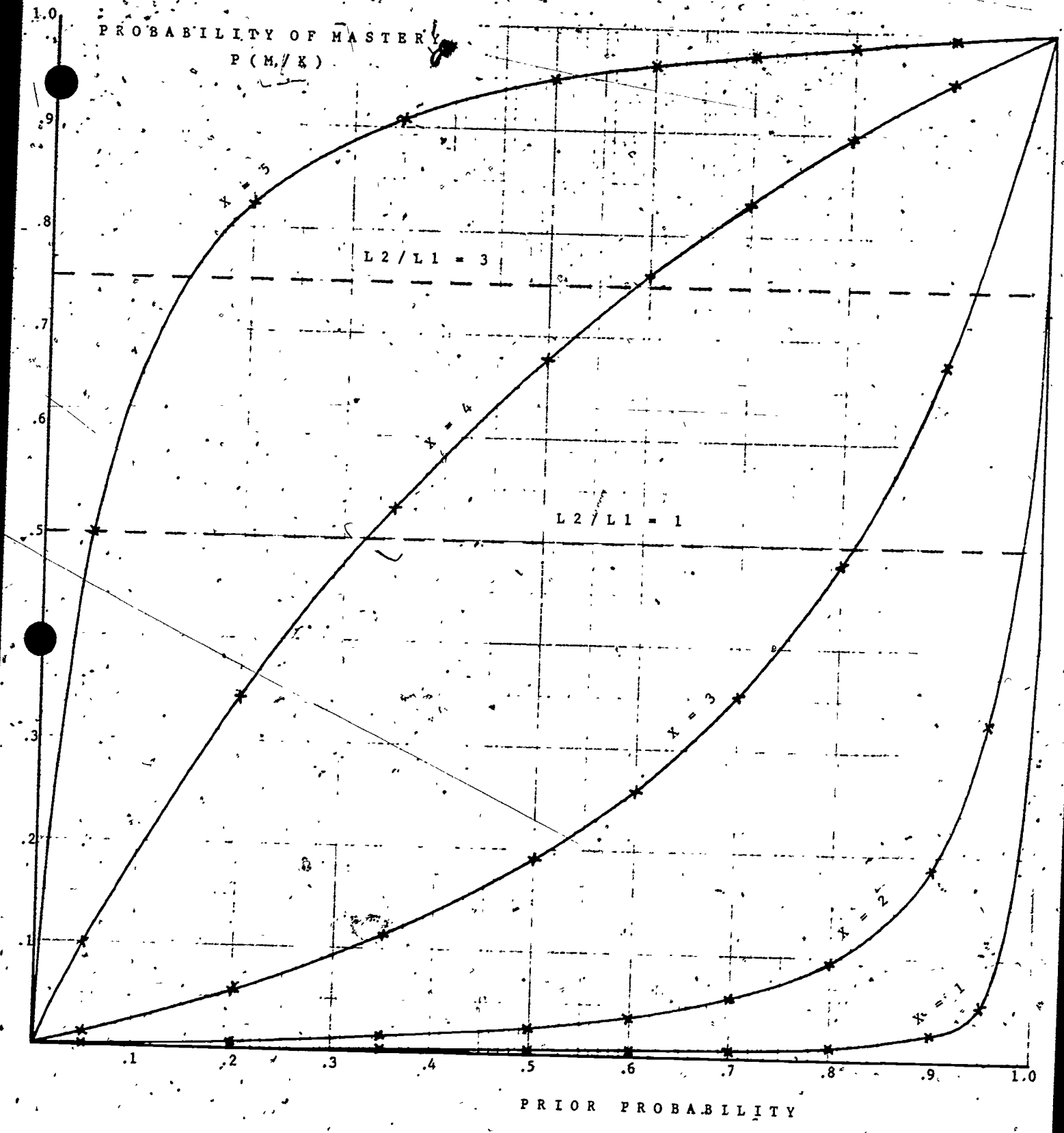

Figure 1. Probability of Mastery as a function of Prior Probability for a 5-item test.
$\alpha = .5$ $\beta = .1$

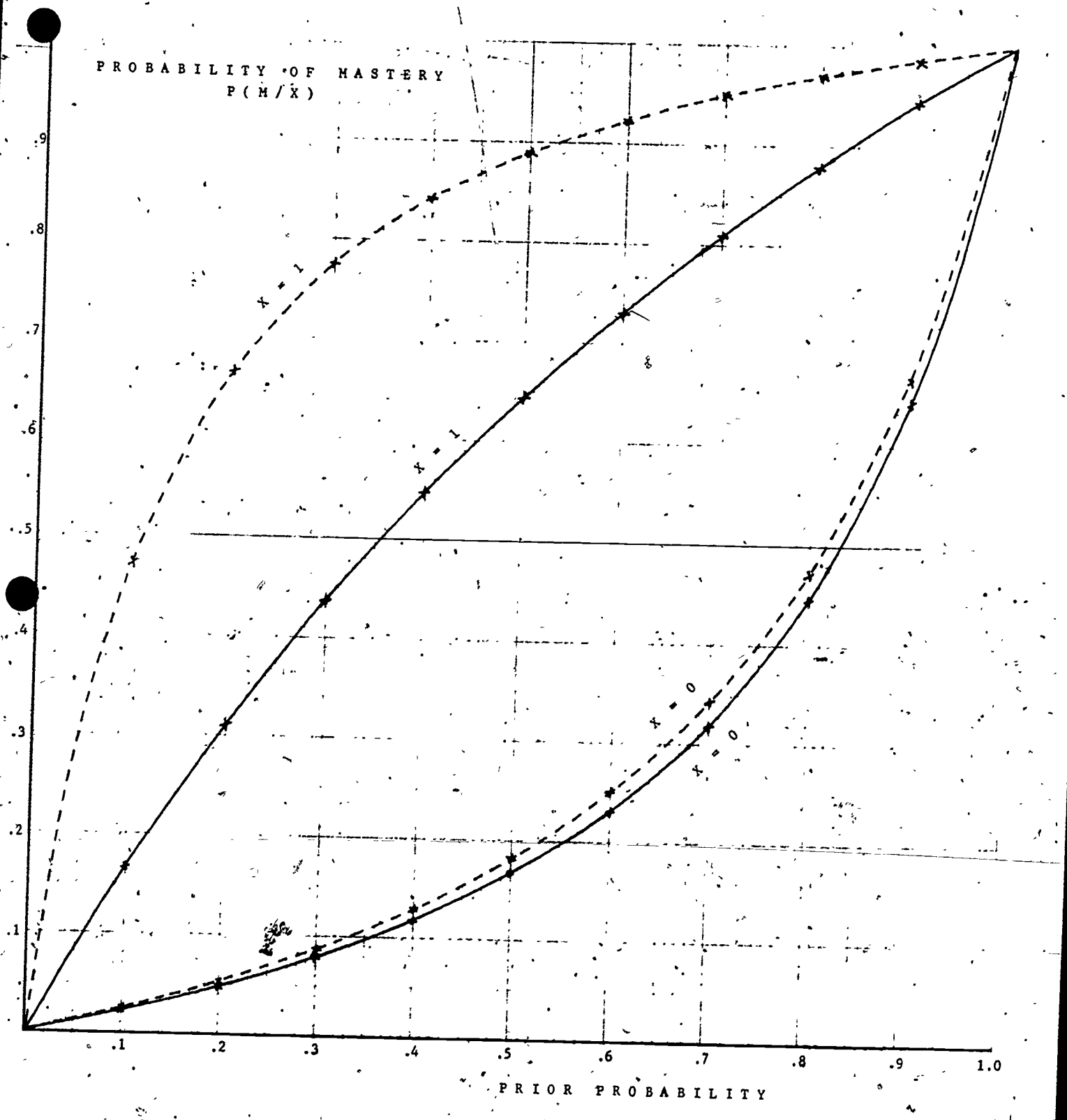

Figure 2. Probability of Mastery as a function of Prior Probability for a 1-item test.
Solid lines: $\alpha = .5$, $\beta = .1$
Dashed lines: $\alpha = .1$, $\beta = .2$